様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １１ 日現在

機関番号：３４５０６
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１５５８

研究課題名（和文）金融テキストマイニング－マーケットセンチメント分析と異言語文書間類似度の推定－

研究課題名（英文）Financial Text Mining: Market Sentiment Analysis and Document Semantic Similarity for Different Languages

研究代表者
関　和広（Seki, Kazuhiro）

甲南大学・知能情報学部・教授

研究者番号：３０４４４５６６

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では，ニュース記事等の大量のテキスト情報を活用し，金融・経済分野におけるテキストマイニングの研究を推進することを目的とし，速報性の高いニュースメディアを基にした景況感予測，および異なる言語で書かれた文書間の類似度の推定に取り組んできた．前者の景況感予測については，近年の自然言語処理タスクで主流となっている自己注意機構を用いたモデルを採用し，頑健かつ精度の高い景況感予測を可能とした．後者の異言語の文書間類似度については，深層学習に基づく翻訳モデルの内部表現を利用することで，従来手法以上の精度で日本語ー英語間，英語ーヒンズー語間での意味的な類似度の推定に成功した．

研究成果の学術的意義や社会的意義
従来はアンケートなどのコスト・時間がかかる集計調査によっていた景況感が低コストかつほぼリアルタイムで行えることが示されたことにより，本研究成果を利用することで，金融当局の政策や企業の意思決定がよりタイムリーかつ効果的に行えるものと期待できる．さらに，異言語の文間の類似度を高精度で推定できることが明らかになったことから，これを発展させ前者と併用することで，言語の壁を超えて金融・経済関連テキストデータを統一的に分析することが可能となる．

研究成果の概要（英文）：In this research, we studied (1) business sentiment forecast and (2) estimation of similarity between documents written in different languages, with the aim of promoting research on text mining in the financial and economic fields by utilizing a large amount of textual information such as news articles. For the former, we employed a model based on a self-attention mechanism, which has become the mainstream in recent natural language processing tasks, and achieved robust and highly accurate business sentiment prediction. For the latter, we successfully estimated the semantic document similarity between Japanese and English, and between English and Hindi, by using the internal representation of a translation model based on deep learning.

研究分野：人工知能

キーワード：深層学習　機械学習　大規模言語モデル　景況感指数　足元予測　多言語モデル　文書間類似度

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

従来のニュースメディアやユーザ発信型の Consumer Generated Media (CGM) など，ウェブ上では膨大なテキスト情報が日々発信されている．通常，これらのテキスト情報は，情報の受け手によって取捨選択され，個人的に消費されることが意図されている．しかしながら，これら大量のテキスト情報を大規模・総合的に分析することで，様々な応用が可能になる．本研究では特に，ウェブ上の既存のニュースメディアを対象にした「経済・金融分野におけるセンチメント分析」，および記述言語をこえたテキストの有効利用を図るための「異言語文書間の類似度推定」に関する研究を実施する．以下，それぞれの背景について簡単に述べる．

(1) 経済・金融分野におけるセンチメント分析

CGM を対象とした主要な研究の一つとして，分析対象のテキストに表現される書き手の感情の抽出や指数化を行う感情分析がある．感情分析には，極性辞書や感情表現辞書がよく使われ，これらの辞書には，肯定や否定，あるいは喜びや悲しみといった感情を表現する語が収録されている．これらの辞書の多くは一般的なドメインを対象にしたものであるが，金融・経済の分野でも極性辞書構築の研究が行われている．従来の感情分析が，文章の書き手の感情を分析するものであるのに対して，金融・経済分野ではマーケットセンチメント（市場感情）やビジネスセンチメント（景況感）を分析の対象とする．マーケットセンチメントとは，特定の銘柄に対する投資家の気分や態度の総体であり，たとえば株価の動きとして表出する．一方，ビジネスセンチメントは，経済やビジネスの状態や見通しに対する人々の感じ方や評価を指す概念であり，特定の時点や期間における経済の健全性やビジネスの現状に対する意識や評価を反映する．

(2) 異言語文書間の類似度推定

ウェブ等に存在する大規模なテキスト情報を有効活用するためには，情報の検索や組織化の技術（クラスタリングなど）が必要である．このようなテキスト処理で用いられる基本的な尺度として，文書間の類似度がある．文書間の類似尺度はさまざまな定義が可能であるが，古典的には，Bag-of-Words（語順を無視した表現方法）によって各文書をベクトルで表現し，ベクトル間の内積やコサイン類似度で測ることが多い．この大前提として，2 つの文書は同一の言語で書かれている必要がある．しかしながら，金融・経済がグローバル化した現代では，日本語などの単一の言語に頼った情報収集や分析では，　単に求める情報を見つけられないだけではなく，情報のバイアスや欠落といった危険性もある．もし，異なる言語で書かれた 2 文書間の類似度を測ることができれば，このような問題が緩和あるいは解消され，異言語の文書群を統一的に分析したり，一つの言語によって他の言語を横断的に検索したりといった応用が可能になる．

２．研究の目的

上述のような背景から，本研究では，ウェブ上の大量のテキスト資源の金融・経済ドメインへの効果的利用を目的として，ニュース記事等を用いた景況感の足元予測，および異言語文間の類似度推定を行う．

３．研究の方法

(1) ニュース記事等を用いた景況感の足元予測

従来の景況感指数として，内閣府が毎月発表している景気ウォッチャー調査がある．本研究では，このデータから景気判断および判断理由文のペアを抽出し，2 つのモデルを学習する．一つはテキストデータから景気判断を行うモデルであり，もう一つは外れ値を検出モデルである．前者については，Bidirectional Encoder Representations from Transformers (BERT)を採用する．具体的には，日本語の大規模なコーパスで事前学習された BER に，景気判断を実数で予測する出力層を加え，景気ウォッチャー調査を学習データとしてモデル全体の重みをファインチューニングする．後者の外れ値検出モデルは，入力テキストのフィルタリングに用いる．本研究で景況感指数の予測に利用するデータはニュース記事であり，ニュースには多様な記事が含まれるため，外れ値検出によって経済や景気に関する文だけを選択的に利用する．外れ値検出のモデルには，1 クラス Support Vector Machine（SVM）を利用する．モデルの学習には景気ウォッチャー調査の景気判断理由文を与え，景気判断理由文と類似している文だけを経済・景気に関係する文と見なして，景況感指数の算出に利用する．より具体的には，ニュース記事を句点「。」をもとに文に分割し，文ごとに外れ値検出モデルに入力する．その結果外れ値と判定された文は除外し，そ

れ以外の文を景気判断モデルに入力する．その結果，文ごとに景気スコアが出力される．出力された景気スコアを月毎にまとめて平均値を算出し，それを月次の景況感指数とする．

なお，景況感は，金融政策や物価，為替，雇用，賃金，海外情勢など，様々な要因から形成されている．しかしながら，全ての要因が経済に等しく影響を与えているわけではなく，経済の情勢判断を行う上では，経済を上向きあるいは下向きに動かす要因が何であるのかを知ることが重要である．そのため，本研究では，上述の方法で推定した景況感指数に対する語の寄与を算出し，どのような要因（語句）が，いつ，どの程度景況感に影響を与えたのかを分析する．

(2) 異言語文間の類似度推定

異言語文間の類似度を算出する簡単な方法は，言語横断検索等で従来行われてきたよう，機械翻訳システムを用いて 2 つの言語を同一言語に翻訳した上で，Bag-of-Words で文書をベクトル表現して類似度を測ることである．しかしながら，この方法には次のような欠点がある．

- 利用する機械翻訳システムの精度（単語レベルの誤訳など）に影響を受ける．
- 文脈や係り受けなど単語間の意味的・統語的依存関係を考慮できない．

本提案研究では，ニューラル機械翻訳モデルを利用することで，上記の問題を回避あるいは低減しつつ，異言語文書間の類似度を算出する．具体的には，Sequence-to-Sequence 機械翻訳（MT）モデルを利用し，エンコーダの内部状態によってそれぞれの言語の文章をベクトルとして表現する．そして，2 言語のベクトル空間の間の翻訳行列を直交制約を加えて学習することで，異言語間のベクトルの類似度を算出する．この方法は翻訳元の言語を（中間表現には変換するが）翻訳先の言語に訳出しないため，上述の誤訳の影響を軽減できると考えられる．また，Sequence-to-Sequence モデルは長短期の時間的な依存関係を扱うことができるため，2 つ目の語の依存関係の問題もある程度解消できるものと期待される．さらに，MT モデルの翻訳結果の n ベスト解を Bag-of-Words でベクトル表現し，もう一つの異言語文間類似度を算出する．最後に，二つの類似度をランキング学習によって統合することで，さらなる精度向上を図る．

４．研究成果

(1) ニュース記事等を用いた景況感の足元予測

前述の外れ値検出モデル，および景気判断モデルを用いて，2008 年 1 月～2020 年 6 月までに発行された日経新聞の記事見出しと記事本文を入力として景況感指数を算出した．アンケートに基づく従来の景況感指数（内閣府発表の景気ウォッチャーDI）と比較したところ，リーマンブラザーズの経営破綻に端を発する金融危機や東日本大震災による景況感の低下など，本景況感指数はおおむね景気ウォッチャーDI のトレンドに近い動きを示しており，実際に両者の相関係数も 0.888 と高い正の相関があった．なお，今回実験に利用したデータは一般の経済・社会情勢を伝える全国紙であり，新聞記事だけを入力として，直接的に景気について調査したアンケートに基づく景気ウォッチャーDI に近い結果が得られたことは特筆すべきであり，本景況感指数の妥当性・有用性を示すものである．

なお，景気ウォッチャー調査の回答者は，約 7 割が家計動向関連，約 2 割が企業動向関連，約 1 割が雇用関連の業種に就いており，家計動向の影響が相対的に大きい．よって，景気ウォッチャーDI も家計動向の影響をより強く受けた指数であると言える．一方，本研究で指数の推定に用いた日経新聞は経済紙であるため，ビジネス関連の記事が多いと考えられる．そこで，企業動向関連業従事者の回答のみから算出された景気ウォッチャーDI と本景況感指数を比較したところ，相関係数は 0.888 から 0.937 に上昇した．この結果は，日経新聞から算出した本景況感指数は，より企業動向を反映した指数であることを示唆している．また，実験によって本景況感指数は景気ウォッチャーDI の一致指数であることが分かった．ただし，本景況感指数の集計は月末に即時終了するため，景気ウォッチャー調査よりも 10 日程度早く公表が可能である．最後に，文単位で予測した景気スコアを単語単位に分割して再集計することで，任意の要因が景況感に与える影響を時間軸に沿って分析できることを示した．

(2) 異言語文間の類似度推定

提案手法を日英の翻訳データで評価したところ，MT モデルの n ベスト解を用いる手法では，n が 2 以上のとき Prec@1 が 0.338 から 0.550 に，MT モデルの中間状態を用いた手法に直交制約を加えた場合，Prec@1 が 0.504 から 0.659 に向上し，非常に効果的であることがわかった．さらに，両手法を組み合わせた場合，Prec@1 は 0.800 まで向上し，高精度な NMT モデル（Google 翻訳）を利用した場合とほぼ同精度であった．なお，本研究で用いた小規模な MT モデルと Google 翻訳の精度を BLEU で比較すると，前者は 0.039，0.177 と大きな開きがあり，もし仮に Google 翻訳

の内部表現を本研究で用いれば，さらに優れた性能が得られると考えられる．

続いて，従来手法との比較として，Word2Vec の重み付き平均，多言語単語埋め込みモデルの重み付き平均，および文埋め込みを用いたニューラルネットワークベースのモデルについて，日英言語横断文検索タスクで実験を行った．その結果，いずれの手法よりも提案手法が高い性能を示すことが確認された．さらに，英語とヒンディー語の言語ペアでも同様の実験を行い，先行研究（ニューラルネットワークベースの XCNN）と比較して，提案手法の有効性（MRR で XCNN が 0.533 に対して提案手法が 0.708）を示した．最後に，提案手法を言語横断ニュース検索に適用し，与えられた別の言語のニュースと意味的に類似したニュース記事を見つけることができることを示した．

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Kazuhiro Seki, Yusuke Ikuta, and Yoichi Matsubayashi	４．巻 59
２．論文標題 News-based business sentiment and its properties as an economic index	５．発行年 2022年
３．雑誌名 Information Processing & Management	６．最初と最後の頁 ―
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.ipm.2021.102795	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 ―

１．著者名 Kazuhiro Seki	４．巻 47
２．論文標題 Cross-lingual text similarity exploiting neural machine translation models	５．発行年 2021年
３．雑誌名 Journal of Information Science	６．最初と最後の頁 404-418
掲載論文のDOI（デジタルオブジェクト識別子） 10.1177/0165551520912676	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

１．著者名 関和広, 生田祐介	４．巻 62
２．論文標題 経済ニュースによる景況感指数の足元予測	５．発行年 2021年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1288-1297
掲載論文のDOI（デジタルオブジェクト識別子） 10.20729/00211101	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

１．著者名 Kazuhiro Seki and Yusuke Ikuta	４．巻 ―
２．論文標題 S-APIR: News-Based Business Sentiment Index	５．発行年 2020年
３．雑誌名 Proceedings of 24th European Conference on Advances in Databases and Information Systems	６．最初と最後の頁 ―
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

1．著者名 Kazuhiro Seki	4．巻 27
2．論文標題 On Cross-Lingual Text Similarity Using Neural Translation Models	5．発行年 2019年
3．雑誌名 Journal of Information Processing	6．最初と最後の頁 315-321
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.27.315	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Kazuhiro Seki and Yusuke Ikuta	4．巻 -
2．論文標題 Estimating Business Sentiment from News Texts	5．発行年 2019年
3．雑誌名 Proceedings of the 2nd IEEE Artificial Intelligence and Knowledge Engineering	6．最初と最後の頁 55-56
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Kazuhiro Seki	4．巻 -
2．論文標題 Exploring Neural Translation Models for Cross-Lingual Text Similarity	5．発行年 2018年
3．雑誌名 Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)	6．最初と最後の頁 1591-1594
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Kazuhiro Seki	4．巻 -
2．論文標題 Turning News Texts into Business Sentiment	5．発行年 2022年
3．雑誌名 Proceedings of the 44th European Conference on Information Retrieval (ECIR 2022)	6．最初と最後の頁 311-315
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　1件／うち国際学会　0件）

1．発表者名 関和広，生田祐介，松林洋一
2．発表標題 ニュース記事に基づく景気指標S-APIRの開発
3．学会等名 第24回人工知能学会金融情報学研究会
4．発表年 2020年

1．発表者名 関和広，生田祐介，松林洋一
2．発表標題 テキストデータを利用した新しい景況感指標の開発と応用
3．学会等名 AI・ビッグデータ経済モデル研究会（招待講演）
4．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関